REDACTED VERSION

November 26, 2025

Via Electronic Filing

The Honorable Susan van Keulen
United States District Court for the Northern District of California
San Jose Courthouse, Courtroom 6 – 4th Floor
280 South First Street
San Jose, CA 95113

Re: ***In re Google Generative AI Copyright Litigation***
Master File Case No. 5:23-cv-03440-EKL-SVK
Consolidated Case No. 5:24-cv-02531-EKL-SVK

Pursuant to Section 7(b) of the Court's Civil and Discovery Referral Matters Standing Order, Plaintiffs and Defendant Google LLC ("Google") respectfully submit this joint statement regarding the dispute over Google's production of Named Plaintiff and class discovery. Fact discovery closes in 79 days, on February 13, 2026. Counsel have met and conferred but remain at impasse.

Plaintiffs' Position. The Court should compel Google to produce documents sufficient to identify, or simply to identify, (1) all Named Plaintiffs' copyrighted works in the datasets, including pirated datasets, copied and used by Google to train the AI models at-issue: GLaM, LaMDA, PaLM, ULM/PaLM 2, Bard, Imagen, and Gemini ("Models"); and (2) all class member copyrighted works used for that purpose.[1] Plaintiffs submitted a methodology that reliably identifies Class Works in Google's training data. ECF No. 266-1. But Plaintiffs should not have had to incur the time and expense of developing that methodology, as the evidence now squarely contradicts Google's argument that it cannot identify which copyrighted works were copied and used in training. Insulating Google from producing this evidence would also unfairly prejudice Plaintiffs, because Google's knowledge and tracking of copyrighted works bears on the willfulness of its ongoing infringement. Forcing Plaintiffs to hunt through Google's "sample" data sets is plainly less efficient.

The evidence shows that Google maintains records of [redacted] used to train several of the Models. *See, e.g.*, ECF No. 282, Google's Class Certification Opposition, at 2-3 (Google [redacted]); ECF No. 253-86, GOOG-AIC-000481984 at -987 ([redacted]); GOOG-AIC-000388450 at -458 ("[redacted]"). It stores the titles of books in the datasets used to train Gemini, and can identify books in its training datasets by title, author, and publisher. These are all methods for identifying Class Works, which would minimize the substantial expenses Plaintiffs continue to incur through their experts, and thus should be compelled. Google's response is heavy on counsel's argument and adjectives, and light on evidence, and it does not contradict these facts.

Plaintiffs requested this information from the outset of the case. *See* Ex. A, RFPs 5, 11, 13, 27, 65, seeking Google's "internal tools used to . . . catalog Training Data" (RFP 5); documents

[1] "Class Works" refers to the copyrighted works, both text and image, ingested and used by Google to train the Models.

concerning the "inclusion of . . . copyrighted works in datasets (RFP 11); "methodologies" for "organizing, curating . . . or annotating Training Data" (RFP 13); and documents that "identify any of the Plaintiffs, including . . . references to their copyrighted works" (RFP 27). The Parties met extensively about these requests. Google falsely continues to deny it can identify copyrighted works in datasets. *See, e.g.*, ECF 240 at 9 ("***Google does not have preexisting tools or ready means to efficiently and reliably determine whether a specific work can be found in training data from a particular source.***") Instead, Google represented that the best evidence of the copyrighted works were sample datasets used to train Google's Models. *See* Google's Response to Pltfs' ROG 2 ("Google proposed that Plaintiffs direct their attention to one exemplary text model and one exemplary image model"). Conditioned on these false representations, Plaintiffs accepted a compromise and agreed to search sample datasets for Class Works themselves. *See, e.g.*, May 20, 2025 S. Teti Email to P. Sampson. Plaintiffs alternatively sought source code, which would provide a mechanism for determining the use of Class Works, the details of the provenance of training data or how Google used the class works for training its Models. RFP 65 (including source code for annotating and organizing datasets); *see also* ECF 140 at 4. Google opposed that request too, representing that ***source code would not identify "how many and which individuals are part of the class" because the works trained on "may be found in the datasets themselves[.]"*** *Id.* at 10. Cumulative evidence calls Google's position into question because metadata in source code can do just that.

In September 2025, Plaintiffs timely issued narrow discovery that sought evidence of Google's methods for tracking and identifying copyrighted works. They seek identification of the "works included in the Recitation Checker, including all Named Plaintiffs Works in Suit" (ROG 18) and for example, the "metadata fields that may be queried, sorted, or otherwise used to identify files made available for use as training in the database" (ROG 14). The Parties reached impasse on these interrogatories. Out of an abundance of caution, Plaintiffs recently propounded targeted requests seeking documents sufficient to identify copyrighted works "used to build, train, and develop the Models" (RFP 80), "contained in each version of the [redacted] used to train any of the Models, including lists maintained by [redacted]" (RFP 81), and "contained in Google's blocklists" (RFP 83), and documents sufficient to identify the "indices, databases, or training datasets used by Google's Recitation Checkers . . . to identify and block regurgitation of copyrighted works used to build and train the Models" (RFP 84).

Google's counsel continues to argue that the "only means of determining what materials Google used to train its AI models is the training data itself." *Infra* at 6. But what Plaintiffs seek is identification of the copyrighted material Google copied and used to train the Models. It is unclear what Google even means by a "final" training set, when there are many iterations of the Models and training is ongoing. Google tracks the copied data and its "provenance," or source, for many reasons: targeted acquisition of data for the Models, experimenting with data-mixes, analyzing model performance, and developing the Models issued to the public. Google also tracks copyrighted works in its datasets because it recognizes its uncontrolled use of copyrighted works creates "legal risk." ECF 253-81 at -506 ("[redacted]"); *id.* at -045 ("[redacted]"). The same or similar methods Google uses for its own purposes can be used to identify Class Works.

1. ***Title/Author/Publisher Tracking:*** Google developed specific tools to identify books in the [redacted]. GOOG-AIC-0008162990.C at -291.C ("[redacted]"). Google has done so to test which data improves the Models. GOOG-AIC-000387189 at -201 ("[redacted].") Andrew Dai, Principal Researcher and Director at Google DeepMind and Gemini Data Area Lead, testified that the Gemini team "[redacted]," Dai Dep. 360:22-361:3, and that Google can identify books in datasets by title, author, and publisher. *Id.* 246:7-15. He admitted that he could "[redacted]" to get "[redacted]" which he did specifically to test which works helped improve models. Dai Dep. 314:16-317:19. He also testified that Google has [redacted], *id.* 105:8-106:22, and that Google maintains [redacted] *Id.* 360:6-12.

2. ***Blocklist:*** Google maintains a "blocklist" of [redacted]. It is accessible to Product Counsel and the developers of the recitation checker. Dai Dep. 144:22-147:8; *see also* GOOG-AIC-000013927 (Google [redacted]); Baldridge Dep. Tr. 160:12-162:20 (testifying Google [redacted]). Google's documents show they similarly developed a blocklist of [redacted]. *See, e.g.*, GOOG-AIC-000505190 at -216 (referencing the development of a "[redacted]"). The list of such blocked material is an "underlying fact" and not subject to privilege. *See Upjohn Co. v. United States*, 449 U.S. 383, 395 (1981); *Sky Valley Ltd. P'ship v. ATX Sky Valley, Ltd.*, 150 F.R.D. 648, 663 (N.D. Cal. 1993) ("[T]he privilege . . . blocks access only to communications, not to underlying information, data, or documents."). That's particularly so when developers use those facts to develop filters that are the structure of the products at issue.

3. ***Recitation Checker:*** The recitation checker [redacted]. GOOG-AIC-000756004 at-004 ("[redacted]."). Google maintains indices of training data for its recitation checker. Dai Dep. Tr. 382:17-383:4 (the recitation checker "contain[s] an index of the training data that goes into Gemini"). The indices used for the recitation checker are used by Google to [redacted] *See* GOOG-AIC-000756004 at -004 ("[redacted]

[redacted].”). Google’s 30(b)(6) representative confirmed that a recitation checker’s index could be used to determine if a piece of any particular book was in a model training data set. Carver Dep. 75:19-23 ([redacted]).

4. ***Lists Maintained By CDA:*** Li Xiao is the Founder of Google’s Core Data Acquisition Team (“CDA”), the team that acquires data to train Google’s models. He testified that the data collected by CDA on behalf of Google DeepMind for training Gemini is maintained by Google in a form that is searchable, sortable, and retrievable. Xiao Dep. 29:20-31:8; 135:12-137:11; 138:20-140:1; 116:4-20 (testifying that he could “[redacted]).

Google uses all of these tools to identify copyrighted works, including specific subsets of Class Works, for its own purposes. Counsel’s argument that Google cannot identify which data is being used to train the Models and where copyrighted data exists in those datasets, is both not credible, and directly contradicted by testimony from Google’s witnesses. [redacted]; indices were created from some list of works. Plaintiffs are not being prescriptive about the how, and just seek the result. “Courts regularly require parties to produce reports from dynamic databases, holding that ‘the technical burden of creating a new dataset for the instant litigation does not excuse production.’” *Apple Inc. v. Samsung Elecs. Co. Ltd.*, 2013 WL 4426512, at *3 (N.D. Cal. Aug. 14, 2013) (cleaned up); *see also Gonzales v. Google, Inc.*, 234 F.R.D. 674, 683 (N.D. Cal. 2006); *In re eBay Seller Antitrust Litig.*, 2009 WL 3613511, at *2 (N.D. Cal. Oct. 28, 2009). Plaintiffs seek an order requiring Google to produce evidence sufficient to identify, or simply to identify: (1) all Named Plaintiffs’ copyrighted works in the datasets, including pirated datasets, copied and used by Google to train the Models, and (2) all class member copyrighted works used for that purpose.

Google's Position: Plaintiffs' latest motion to compel—the 13th such effort in this litigation—is perhaps their most meritless, and should meet the same fate as the overwhelming majority of its predecessors: outright denial.

The starting, and what should be the ending, point for this motion is that Google does not have, and cannot do, what Plaintiffs demand. There is no list of works Google used for training, and no way for Google to generate one. ***The only record and the only means of determining what materials Google used to train its AI models is the training data itself.*** That is why, for the past nine months, Google has undertaken enormous efforts and spent enormous resources to gather that data and then create and staff a one-of-a-kind data review environment for Plaintiffs. It is also the reason that drove the Court-approved party compromise through which Plaintiffs selected and for months have analyzed representative datasets totaling more than 1.5 petabytes of information. *See* ECF No. 272 at 1; ECF No. 155; ECF No. 161 at 25:23-31:24. After all that, Plaintiffs apparently are unable to do what they repeatedly told the Court they could do (and claim in this and their certification motion that they have done)—identify the specific copyrighted works that were actually used within the gargantuan training datasets (which include filtered webpages taken from Google's automated crawl of the internet).[2] But Plaintiffs' difficulties are no excuse for them to peddle the fabricated narrative that Google can do it after all.

If Plaintiffs have harbored the mistaken belief that Google could somehow solve their intractable problem of identifying all "relevant works" and putative class members, the time for this motion was not now—after Plaintiffs filed their twice-delayed class certification motion—but more than a year ago, before launching a massive discovery campaign. Instead, after moving to certify classes whose members they plainly cannot identify, Plaintiffs served on November 17,

[2] In Plaintiffs' certification motion, they proffered a "data-matching" expert, Meredith McCarron, who purports to have identified a "straightforward," "replicable, consistent, and accurate" methodology for identifying "registered copyrighted" works in the datasets. ECF No. 253-4. If she did, Plaintiffs would not be demanding Google to perform the same task. In reality, McCarron utterly failed, and her analysis is unreliable. *See* ECF No. 295 (Google's motion to exclude McCarron's testimony). The incongruity of demanding what Plaintiffs claim to have already done, and their admission in an earlier iteration of this motion that McCarron did little more than "hunt and peck" in the training datasets, speak for themselves.

2025, vexatious document requests (seeking documents sufficient to identify all class works in training data, when no such documents exist), declared an "impasse" two days later on November 19, long before Google's responses are due, and now are rushing back to Court. Plaintiffs should again be admonished for abusing the discovery process. But even if Plaintiffs had properly presented this issue, Google would be giving them the same answer: It does not have the information Plaintiffs demand, and even attempting to generate that information would impose extraordinary and unjustifiable burdens.

Of importance, Plaintiffs obfuscate the difference between identifying some specific content within the training data, and matching that content to what Plaintiffs call "class works," *supra* n.1—*viz.*, particular works of authorship with specific registered copyrights that are unlicensed, as required for the new classes they proposed for the first time in their class motion.[3] Even if Plaintiffs somehow compiled an accurate list of every work present in years' worth of training data, they would still be nowhere close to justifying certification of these classes. As Google showed in its certification opposition (ECF No. 282), to decide whether any work qualifies as a "class work," one would need to evaluate *inter alia* for each work: whether it is copyrightable; if so, who owns the copyright, whether it is registered, whether the registration is valid, whether registration was within five years of the work's first publication, and whether Google was licensed to use it. Resolving these issues necessarily requires intensely individualized factual investigations and adjudications that cannot be conducted at scale (and that is before addressing individualized issues in multiple affirmative defenses). These have always been Plaintiffs' foremost problem in seeking class certification. As challenging as identification of individual works in training data has been for Plaintiffs and would be for Google, identifying "class works" is even more so.

Plaintiffs Are Misrepresenting the Record. Plaintiffs' motion strains the limits of the joint

[3] Plaintiffs' certification motion proposes new classes that bear no resemblance to the one they pleaded in the operative complaint three weeks earlier. Plaintiffs did not disclose the new classes in discovery, in case management submissions, or in the required meet and confer (which Plaintiffs' failed entirely to conduct). Indeed, Plaintiffs appear to have actively concealed from Google their new, but not improved, proposed classes for months. Based on Plaintiffs' misconduct, Google has moved for sanctions seeking to strike Plaintiffs' class allegations under Rules 16 and 37. *See* ECF No. 298.

discovery letter process. It is not possible to correct each misquotation and fabrication that Plaintiffs offer from deposition and document fragments. Their core misrepresentation, however, is straightforward: No witness, document, or technical record supports the premise that Google maintains, or could somehow readily generate, a complete catalog of *works* with accurate identifying information, let alone a catalog of copyrighted, timely registered, and unlicensed "class works" (*supra* n.1), used for model training. The Court can review the evidence Plaintiffs cite for itself and see the liberty Plaintiffs take with it. And each time Plaintiffs assert a proposition for which no evidence is cited, it is invariably a "tell" that they are making up that proposition. Google briefly describes the actual record below.

[redacted] *Is Not Identifying Training Data or "Class Works."* While Google maintains records of books that were [redacted], that does not identify the training set. Knowing [redacted]—*i.e.*, "How to find out what we have in the corpus," *supra* at 3—does not reveal what survived the steps of preprocessing, filtering, deduplication, and construction of final training datasets, which are significantly different from what was included in the source corpus. The document and deposition snippets Plaintiffs cite concern attributes of the [redacted], not the materials Google ultimately selected from that corpus and used for training. They also in no way suggest any possibility of matching books against the criteria for "Class Works" (such as current, valid copyright registration) that Plaintiffs just added to their proposed class definitions. Again, the only way to determine which works actually appear in the training data is to analyze the training datasets directly, the approach Plaintiffs forced Google to assist them with for months. And the only way to determine whether a work in training data qualifies as a "Class Work" is to conduct a case-by-case, individualized analysis that neither Plaintiffs nor Google has any way of doing.

"Blocklists" and Recitation Checker Cannot Identify "Class Works." Plaintiffs next argue that Google maintains a "blocklist" of copyrighted books and images. Again, that is false.

Google's recitation checker, what Plaintiffs call a "blocklist," is not a list of copyrighted works. For text, it consists of [redacted]. For images, [redacted]. These [redacted] are irreversible; they cannot be "decoded" to reveal the text or images in training data from which they were generated. They simply enable a specific and efficient test: The proposed output of a model is similarly processed as [redacted] and compared against [redacted]. *See* ECF No. 289-2, Decl. of Brian Carver ¶¶ 3-5. In other words, the recitation checker can tell [redacted]. But to use it to check for the presence of class works in the actual training data, Plaintiffs would need to test, one-by-one, every image ever created or every short segment of text ever written. *Id.* ¶¶ 10-15. Plaintiffs do not (and cannot) explain how that would be manageable (or better than analyzing the training data itself). Further, neither the training data nor the recitation checker can identify whether any training material or matched material is a "class work"—*i.e.*, whether it is copyrighted, who owns it, whether/when it was registered, and whether it is licensed. *Id.* ¶ 11.

Google's corporate designee on the recitation checker testified at length, ***five months ago***, about its capabilities and, more importantly, its limitations. Plaintiffs know their description of the tool does not match the evidence.

<u>*CDA Ingestion: Factually Wrong, Legally Irrelevant.*</u> Finally, Plaintiffs misrepresent testimony from Dr. Li Xiao to claim that Google's CDA team maintains a "document" showing "[redacted] that CDA ingested to train," and that this somehow enables Google "to identify copyrighted works … for its own purposes." *Supra* at 5. Again, that is false.

CDA's records do ***not*** show what works were used to train any model. Dr. Xiao expressly testified that he does not know what data acquired through Google's web crawls was incorporated into any training dataset. *See, e.g.*, ECF 237-3, Xiao Dep. Tr. at 151:1-7 (testifying he "do[es] not know" what datasets are used for training). And information on what Google merely "ingested" through those crawls is not germane to this case. Indeed, just two weeks ago this Court rejected Plaintiffs' discovery demand predicated on an untimely attempted shift in their case from one challenging training on works to one addressed to "mere 'ingestion'" of works. ECF No. 272 at 3.

It also bears repeating that if Google actually had lists of specific works captured in its web-crawling (and it does not), that would say nothing about which of those works were used in training, or the copyright status of such works, their ownership, registration, or licensing, which is what Plaintiffs need to show just to begin to identify "class works" and potential class members.

Ultimately, Plaintiffs can point to no evidence—no document or deposition testimony—showing that Google can conjure an accurate and complete list of works, much less timely-registered, copyrighted, and unlicensed works, actually used to train the at-issue models. If such evidence existed, surely Plaintiffs would cite it. But it does not exist because the claim that Google can do what Plaintiffs demand is untrue. Google has produced the datasets that would need to be examined to identify the works used to train its models, which included all existing metadata in those datasets (e.g., URLs or author and title information where that metadata is present in the training data). Plaintiffs' problem is that even with the relevant information they cannot identify the works used, let alone the relevant works for the classes they propose to represent.

Conclusion: It is damning that Plaintiffs only began pursuing this discovery and motion a month after filing their class motion and boasting to the Court that they can readily identify a class of copyright holders whose works Google used for training. If Plaintiffs could do (or did) as they boasted, their demand here would be redundant. Regardless, what they now demand of Google is not feasible.

The training datasets are the only materials that identify what data Google used for training. Plaintiff got that information long ago through an agreed-upon, Court-approved, review protocol. Plaintiffs' mischaracterizations of [redacted], technical safeguards, and crawl logs as a catalog of "class works" goes beyond fair advocacy. Their motion must be denied.

Respectfully submitted,

Dated: November 26, 2025

By: */s/ Lesley E. Weaver*

Lesley E. Weaver (SBN 191305)
Anne K. Davis (SBN 267909)
Joshua D. Samra (SBN 313050)
BLEICHMAR FONTI & AULD LLP
1330 Broadway, Suite 630
Oakland, CA 94612
Telephone: (415) 445-4003
lweaver@bfalaw.com
adavis@bfalaw.com
jsamra@bfalaw.com

Gregory S. Mullens (admitted *pro hac vice*)
BLEICHMAR FONTI & AULD LLP
75 Virginia Road, 2nd Floor
White Plains, NY 10603
Telephone: (415) 445-4006
gmullens@bfalaw.com

Joseph R. Saveri (SBN 130064)
Cadio Zirpoli (SBN 179108)
Christopher K.L. Young (SBN 318371)
Evan A. Creutz (SBN 349728)
Elissa A. Buchanan (SBN 249996)
Aaron Cera (SBN 351163)
Louis Kessler (SBN 243703)
Alexander Zeng (SBN 360220)
JOSEPH SAVERI LAW FIRM, LLP
601 California Street, Suite 1505
San Francisco, CA 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
jsaveri@saverilawfirm.com
czirpoli@saverilawfirm.com
cyoung@saverilawfirm.com
ecreutz@saverilawfirm.com
eabuchanan@saverilawfirm.com
acera@saverilawfirm.com
lkessler@saverilawfirm.com
azeng@saverilawfirm.com

Ryan J. Clarkson (SBN 257074)
Yana Hart (SBN 306499)
Mark I. Richards (SBN 321252)
CLARKSON LAW FIRM, P.C.

22525 Pacific Coast Highway
Malibu, CA 90265
Telephone: 213-788-4050
rclarkson@clarksonlawfirm.com
yhart@clarksonlawfirm.com
mrichards@clarksonlawfirm.com

Tracey Cowan (SBN 250053)
CLARKSON LAW FIRM, P.C.
95 Third Street, Second Floor
San Francisco, CA 94103
Telephone: (213) 788-4050
tcowan@clarksonlawfirm.com

Brian D. Clark (admitted *pro hac vice*)
Laura M. Matson (admitted *pro hac vice*)
Arielle S. Wagner (admitted *pro hac vice*)
Consuela Abotsi-Kowu (admitted *pro hac vice*)
LOCKRIDGE GRINDAL NAUEN PLLP
100 Washington Avenue South, Suite 2200
Minneapolis, MN 55401
Telephone: (612) 339-6900
Facsimile: (612) 339-0981
bdclark@locklaw.com
lmmatson@locklaw.com
aswagner@locklaw.com
cmabotsi-kowo@locklaw.com

Stephen J. Teti (admitted *pro hac vice*)
LOCKRIDGE GRINDAL NAUEN PLLP
265 Franklin Street, Suite 1702
Boston, MA 02110
Telephone: (617) 456-7701
sjteti@locklaw.com

Counsel for Individual and Representative Plaintiffs and the Proposed Class

Dated: November 26, 2025 By: */s/ Paul J. Sampson*

Paul J. Sampson
WILSON SONSINI GOODRICH & ROSATI, P.C.
95 S State Street, Suite 1000
Salt Lake City, UT 84111
Telephone: (801) 401-8510
Email: psampson@wsgr.com

David H. Kramer
Maura L. Rees
Qifan Huan
Kelly M. Knoll
WILSON SONSINI GOODRICH & ROSATI, P.C.
650 Page Mill Road
Palo Alto, CA 94304-1050
Telephone: (650) 493-9300
Fax: (650) 565-5100
Email: dkramer@wsgr.com
Email: mrees@wsgr.com
Email: qhuan@wsgr.com
Email: kknoll@wsgr.com

Eric P. Tuttle
Madison Welsh
WILSON SONSINI GOODRICH & ROSATI, P.C.
701 Fifth Avenue, Suite 5100
Seattle, WA 98104-7036
Telephone: (206) 883-2500
Fax: (206) 883-2699
Email: eric.tuttle@wsgr.com
Email: mjwelsh@wsgr.com

Jeremy Paul Auster
WILSON SONSINI GOODRICH & ROSATI, P.C.
1301 6th Avenue
New York, NY 10019
212-453-2862
Email: jauster@wsgr.com

Counsel for Defendant Google LLC

SIGNATURE ATTESTATION

I, Lesley E. Weaver, am the ECF User whose ID and password are being used to file this document. In compliance with N.D. Cal. Civil L.R. 5-1(i)(3), I hereby attest that the concurrence in the filing of this document has been obtained from the other signatory.

Dated: November 26, 2025

/s/ Lesley E. Weaver
Lesley E. Weaver